

Determining Weights of the Attributes in the Model Generalization of River Networks

Alper Sen and Turkey Gokgoz

Yildiz Technical University, Geomatic Engineering Dept., Istanbul, Turkey

Abstract. Finding appropriate weight of each attribute is one of the main points in the model generalization via clustering and classification methods, giving more or less importance to the geographic information. The purpose of this study is to determine the weights of the geometric, topologic and semantic attributes by implementing a statistical method which is chi-squared test of independence in order to calculate their contributions to the overall goal of selection. Two different drainage patterns in the USGS National Hydrographic Datasets which are dendritic and trellis with rectangular at 1:24,000-scale were weighted.

Keywords: Model Generalization, Chi-squared, Weighting, River Network

1. Introduction

Generalization is a process used for reducing the volume of data of a spatial data set while preserving important structures (Sester 2008). Map generalization operations could be grouped into two categories: (1) model generalization which is a filtering process to obtain a subset of a target database for data analysis (Joao 1998), and (2) cartographic generalization which is the set of operations concerned with the optimal visualization of the selected data (Mackaness 2007). Selection is often used interchangeably with the model generalization and database abstraction. Töpfer's Radical Law (Töpfer and Pillewiser 1966) is the only quantitative rule in the selection of the features. It yields the number of features to be displayed, but does not reveal which of the features should be chosen.

In the river network generalization, two questions should be answered: How many branches to be selected? Which branches are important? Töpfer's Radical Law has answered the first question. But second question is not easy to answer.

With reference to United States Geological Survey (USGS) Draft Standards for 1:100,000 National Hydrography Dataset (NHD), capture conditions are

- If stream/river is perennial and flows from lake/pond or spring/seep, then capture.
- If stream/river is intermittent, and can be definitely located, and flows from lake/pond or spring/seep, then capture.
- If stream/river is perennial and is greater than or equal to 0.63" along the longest axis, then capture.
- If stream/river is intermittent, and can be definitely located, and is not in an arid region, and is greater than or equal to 0.63" along the longest axis, then capture.
- If stream/river is intermittent, and can be definitely located, and is in an arid region, and greater than or equal to 1.2" along the longest axis, then capture.

If the whole context of river network is not considered with geometric, topologic and semantic attributes, the abstraction will dramatically destroy the original structure. "Which attributes contribute to the selection of rivers?" and "Comparing the attributes to the original river networks, which one is more important in model generalization?" These questions are very important in the decision of selection in terms of preserving the original structure of river network. Finding appropriate weight of each attribute is one of the main points in model generalization via clustering and classification, giving more or less importance to the geographic information.

In the generalization literature related to weighting, Bjorke (1997) showed how a weight function can be used to control the spatial distribution of the map symbols. Wolf (1988) suggested a weighted network data for the hydrographic generalization. Jiang and Harrie (2004) were empirically weighted their attributes to be used in the road network generalization by Self Organizing Maps. Zhou and Jones (2005) introduced weighted effective area, a set of area-based metrics for cartographic line generalization. Kulik et al. (2005) assigned weights consist of geometric and semantic components to each vertex of every line for deleting points in line simplification. Podolskaya (2007) used weights in quality assessment of generalization. Gulgen and Gokgoz (2011) used the weighted mean calculated by using the inverse of the number of roads with the same connectivity value for the threshold value determining the important roads.

Lotfi and Fallahnejad (2010) categorized various methods for finding weights in the multi attribute decision making literature into two groups:

subjective and objective weights. Subjective weights are determined only according to the preference decision makers. The AHP method, weighted least squares method and Delphi method belong in this category. The objective methods determine weights by solving mathematical models without any consideration of the decision maker's preferences, for example, the entropy method, multiple objective programming, principal element analysis, etc.

It may be a useful way to compare different datasets at different resolutions to determine the contribution of the geometric, topologic and semantic attributes to the overall goal of selection. To this end, in this study, weights were determined by implementing the chi-square (χ^2) test of independence to find the association between attributes and selection, and Cramer's V coefficient (ϕ_c) (Cramer 1946) to find the strength of association. Each numerical weight is calculated by normalization dividing each ϕ_c with the sum of ϕ_c s. Two different drainage patterns in the USGS NHDs which are dendritic and trellis with rectangular at 1:24,000-scale were weighted.

2. Weighting by Chi-Squared Test of Independence

Pearson (1900)'s χ^2 test is a very popular non-parametric test and used for multinomial data. Hence, the values of a quantitative variable must be transformed to the classes of categorical variable (LeBlanc 2004).

Pearson's χ^2 statistic is

$$T = \sum (X - E)^2 / E \quad (1)$$

where E is the expected frequency of the observed frequency X . If T is greater than the critical value which has significance level α and degree of freedom k , H_0 is rejected and H_1 is accepted. Figure 1 shows the right tailed χ^2 curve, and acceptance and rejection regions at significance level α .

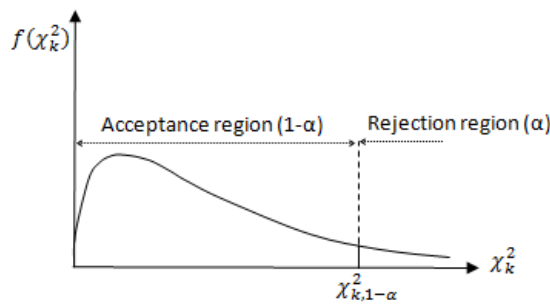


Figure 1. Right tailed χ^2 curve with acceptance and rejection regions at significance level α .

The χ^2 test of independence (also called contingency table analysis) is used to evaluate whether or not there is an association between individuals falling into specific classes of one categorical variable and their membership in classes of a second categorical variable. If we reject the null hypothesis (H_0) of independence, this is equivalent of concluding that there is an association (LeBlanc 2004).

In this study,

- H_0 : *There is no association between an attribute and the selection result.*
- H_1 : *There is an association between an attribute and the selection result.*

In order to weight the attributes, strength of χ^2 association is calculated by ϕ_c (equation 2) and normalized as the sum of weights is equal to one.

$$\phi_c = \sqrt{T/n(m-1)} \quad (2)$$

where n is the number of objects and m is the smaller number of classes in two categorical variables compared.

3. Case Study

In this study, we used four NHDs which are vector geospatial data layers of the National Map, being developed by the USGS. They include two different drainage patterns at 1:24,000-scale and 1:100,000-scale. Pomme De Terre (PT) shows dendritic pattern, while South Branch Potomac (SBP) shows trellis with rectangular pattern. The properties of drainage patterns are described in Howard (1967) and Debarry (2004). Sample drainage patterns are shown in Figure 2.

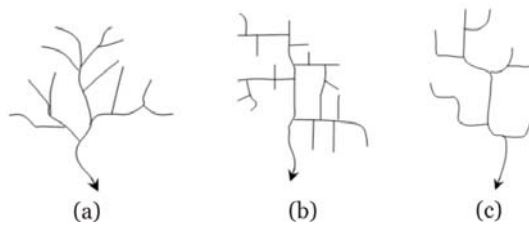


Figure 2. (a) Dendritic, (b) trellis, and (c) rectangular pattern (Debarry 2004).

From the view of selection, it is important that a river should be processed as a whole, and segments should not be eliminated separately, which may disconnect the graph (Stanislawski and Savino 2011). Thus, the river segments were combined using the data of stream level and river type, which

are embedded in the datasets. Following seven geometric, topologic and semantic attributes were considered as associated with the selection of the river networks. The NHD database was enriched with some new attributes as follows.

1. Geometric attributes:
 - a) Length (ratio-scaled): The lengths of the rivers.
 - b) Sinuosity (ratio-scaled): The ratio of the Euclidean distance between the end points to the river length.
2. Topologic attributes: The topologic attributes of the river networks were calculated by three popular centrality measures. The idea of centrality is specifically concerned with communication in small groups and a relationship between structural centrality and influence in group processes (Freeman, 1978).
 - a) Degree centrality (ratio-scaled): Number of links that a node has.
 - b) Betweenness centrality (ratio-scaled): Frequency of shortest linking between the nodes p_i and p_j that node p_k resides on.
 - c) Closeness (ratio-scaled): Inverse of the distance of each node to every other node in the network.
3. Semantic attributes:
 - a) Stream level: A stream level in NHD is a numeric code that identifies a hierarchy of main paths of water flow through the network. Level values are established for the purpose of computationally traversing the drainage network through flow relations identified between the reaches which are the segments of a river network.
 - b) River type: There are two river types. Perennial rivers are those that flow continuously, whereas intermittent rivers appear to dry up when the flow has the potential of being totally absorbed by the bed and underlying material. Intermittent rivers may flow continuously during wet years ("0" and "1" values were assigned to intermittent and perennial rivers, respectively).

The Kolmogorov-Smirnov test for goodness-of-fit (Massey, 1951) was applied for determining the normality of the attributes. All attributes have skewed distributions. Thus, we need a non-parametric test for testing the independence. The χ^2 test of independence is very useful for this case.

Because of using for multinomial data, the length attribute was categorized into two classes regarding to the USGS Draft Standards for 1:100,000 NHD: If a river is longer than or equal to 1.6km, then it is included into the

first, otherwise into the second class. The other ratio-scaled attributes were categorized based on their standard deviations (1σ interval).

In this study, for calculating the weight of each parameter in terms of its contribution to the overall goal of selection, the associations between an attribute and selection result assigned binary integers (0: eliminated; 1: selected) at original were compared. If there is an association, H_o hypothesis (there is no association) is rejected and the difference between observed and expected frequencies is significant. The ϕ_c provided the strength of association (equation 2). The determined weights were calculated by normalizing as the sum of weights is equal to one (dividing each ϕ_c by the sum of ϕ_{cs}).

4. Results

The determined weights of the attributes are given in Table 1 and Figure 3. They are determined with respect to three significance levels (α : 0.05, 0.01 and 0.001).

Subbasin	Length	Sinuosity	Degree	Between	Closeness	Str. Level	Type
PT	30%	14%	19%	21%	No association	11%	5%
SBP	21%	7%	18%	12%	8%	14%	20%

Table 1. The determined weights of the attributes for PT and SBP.

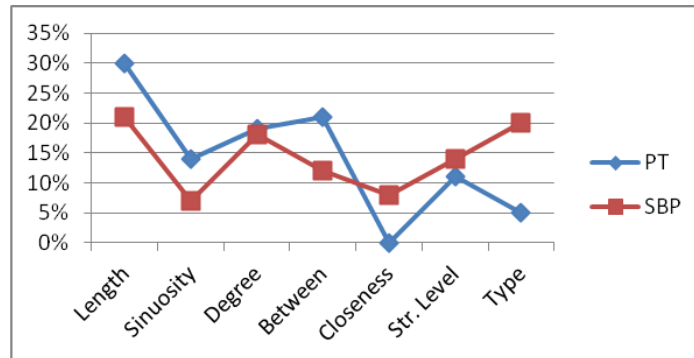


Figure 3. The graphic of the determined weights of the attributes for PT and SBP.

In Figure 4a and 5a, drainage networks of PT and SBP at 1:100,000-scales are shown at 1:1,000,000-scale and 1:1,500,000-scale, respectively. In the close-ups (i.e. Figure 4b-h and 5b-h), drainage networks at 1:100,000 and 1:24,000-scale are shown with together. The branches with the buffers are that of the drainage networks at 1:100,000-scale. Moreover, each branch is

colored with respect to the classes of the attributes as they are in the legends. Furthermore, the black lines in the close-ups are the centerlines. The results for each attributes can be summarized as follows.

- The weights indicate that the length is the most important attribute associated with selection. As they are shown in Figure 4b and 5b, the rivers in red (≥ 1.6 km) are mostly selected for 1:100,000-scale. However, it is needed to the rivers in yellow (< 1.6 km) in behalf of drainage continuity.
- Generally, very sinuous rivers are selected for 1:100,000-scale (red and yellow in Figure 4e and 5h). However, the sinuosity does not guarantee the continuity as well.
- The degree centrality gives the main rivers which have more links (Figure 4d and 5d). However, it could not provide the continuity. The weights of degree centrality are very close for both river networks.
- The betweenness centrality could give both main rivers and linkages between centerlines by small river segments. It is more effective in PT (Figure 4c and 5f).
- The closeness centrality of PT is not associated with the selection result because the χ^2 statistic of closeness is smaller than the critical value. As a result, in Figure 4h, almost all rivers are in green, i.e. there is only one class. However, in Figure 5g, almost all rivers are not in green, i.e. there more than one class, but they reflect weak association. To be honest, this attribute is not so useful for selection.
- The stream level could show the main rivers and provide linkage between centerlines with high hierarchy values. However, it is not so powerful for middle levels (Figure 4f and 5e).
- The weight of river type is higher in SBP, because the perennial rivers are homogeneously distributed, and connection between the main rivers and the intermittent rivers in the trellis with the rectangular pattern of SBP are the perennial rivers. On the other hand, the perennial rivers are heterogeneously distributed in the dendritic pattern of PT (Figure 4g and 5c).

5. Conclusions

Weighting of the attributes is useful in model generalization via clustering or classification. The proposed approach could be used for this aim. Note that, the weights should not be perceived in general, just only specific for PT and SBP. Generic weights may be determined with the weights to be calculated in some more characteristic subbasins conterminous US regarding all patterns.

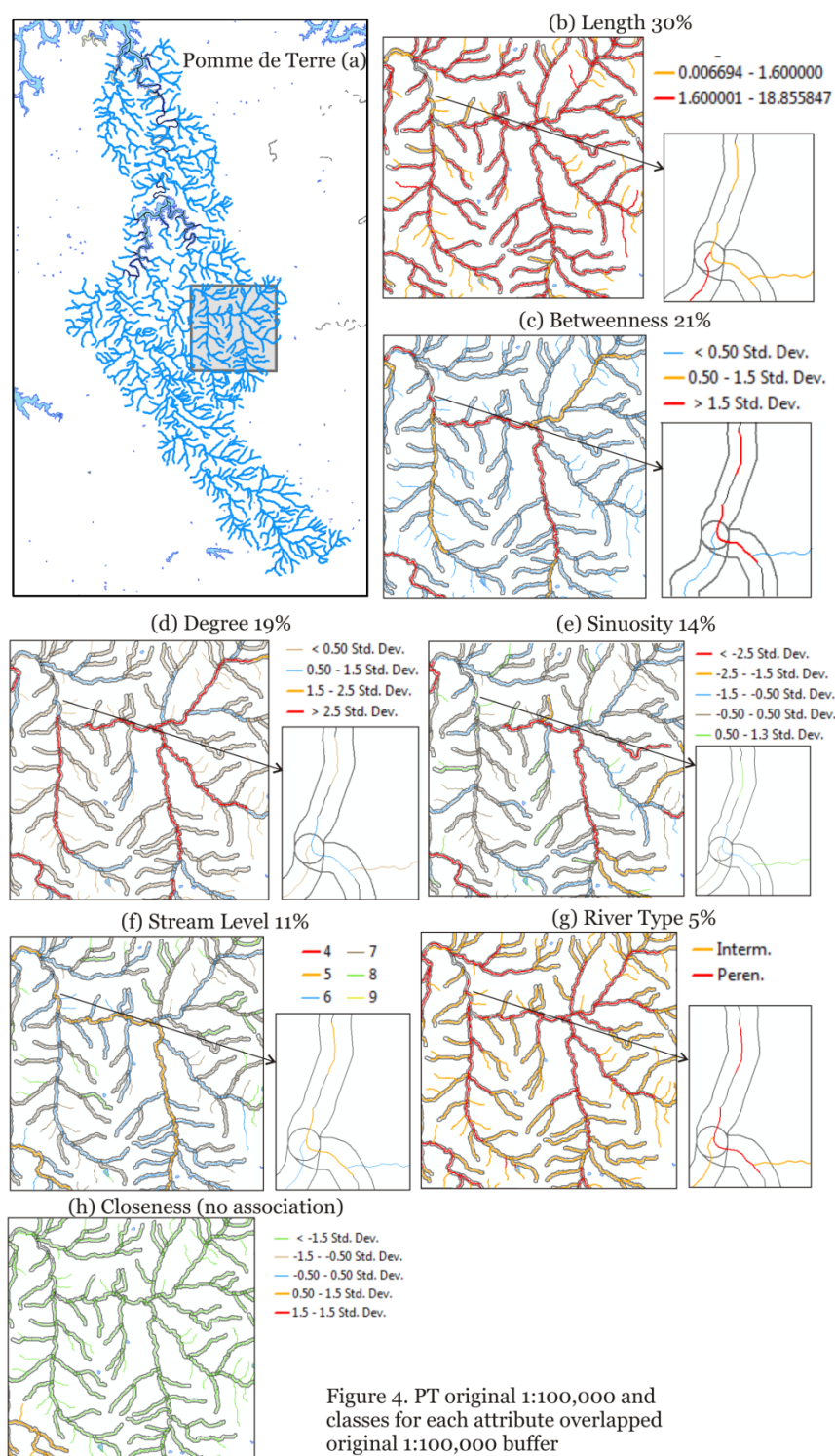
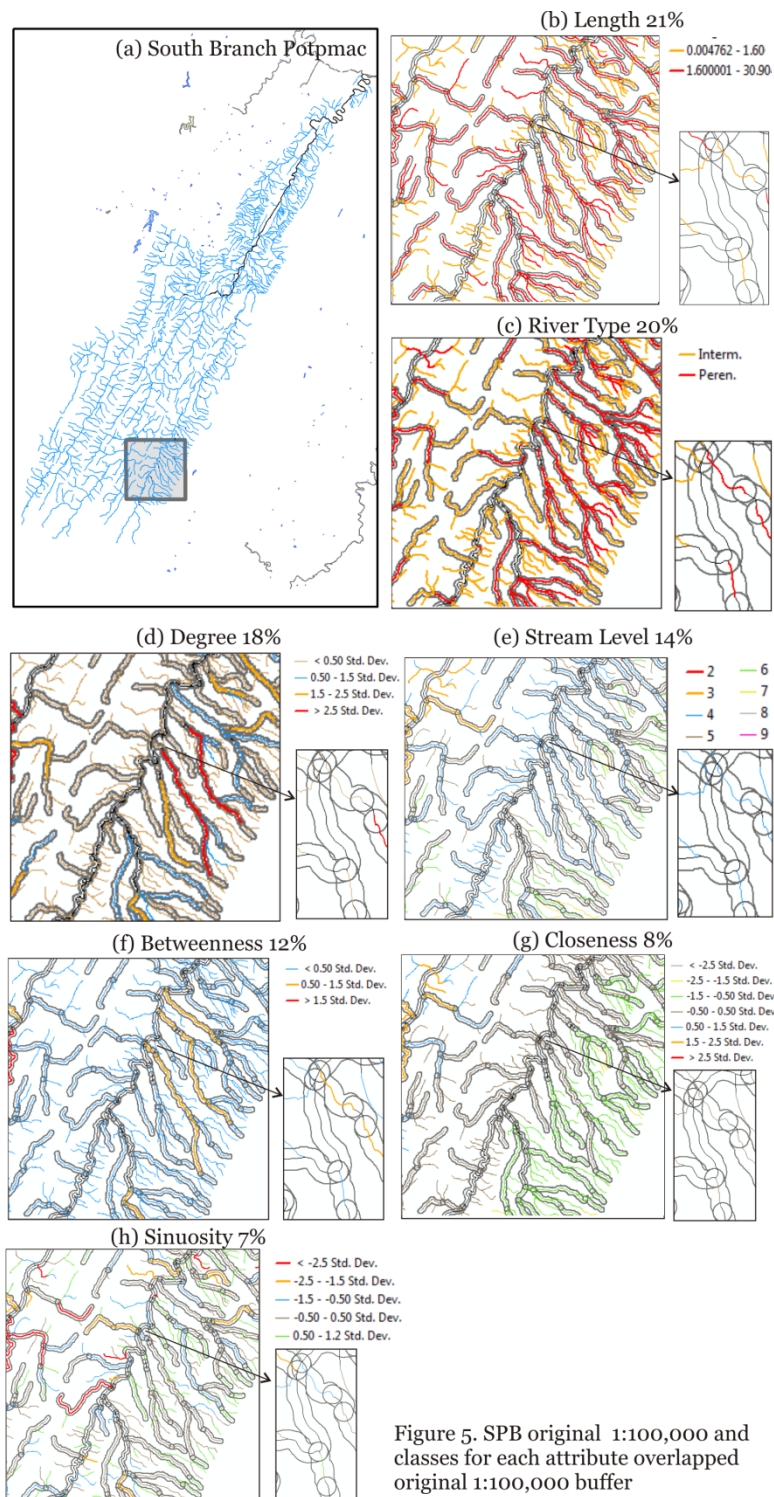


Figure 4. PT original 1:100,000 and classes for each attribute overlapped original 1:100,000 buffer



References

- Björke JT (1997) Map generalization: an information theoretic approach to feature elimination. In proceedings 18th ICA/ACI International Cartographic Conference, edited by Ottoson L. Volume 1, Gavle, June 23-27, Swedish Cartographic Society: 480-486.
- Cramér H (1946) *Mathematical Methods of Statistics*. Princeton: Princeton University Press, 282.
- DeBarry PA (2004) *Watersheds: Processes, assessment, and management*. New Jersey: Wiley.
- Freeman LC (1978) Centrality in social networks conceptual clarification. *Social Networks*. (1) Netherlands: Elsevier, 215-239.
- Gülgen F, Gökğöz T (2011) A Block-based selection method for road network generalization. *International Journal of Digital Earth*, 4(2): 133-153.
- Howard AD (1967) Drainage Analysis in Geologic Interpretation a Summation. *The American Association of Petroleum Geologists Bulletin*. 51:11.
- Jiang B, Harrie L (2004) Selection of streets from a network using self-organizing maps. *Transactions in GIS*. 8(3): 35-350.
- Joao EM (1998) *Causes and Consequences of Map Generalisation*. Taylor & Francis Ltd. London: 35-37.
- Kulik I, Duckham M, Egenhofer, MJ (2005) Ontology-driven map generalization. *Journal of Visual Languages and Computing* 16(2).
- LeBlanc, D. 2004. *Statistics: Concepts and Applications for Science*. Jones & Bartlett Learning.
- Lotfi FH, Fallahnejad R (2010) Imprecise Shannon's Entropy and Multi Attribute Decision Making. *Entropy* 12, 53-62; doi:10.3390/e12010053
- Mackaness WA (2007) Understanding geographic space. In *Generalization of Geographic Information: Cartographic Modelling and Applications*, edited by Mackaness WA, Ruas A, Sarjakoski LT. Amsterdam: Elsevier 1-10.
- Massey FJ (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*. Vol. 46, No. 253, 68-78.
- Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5* 50 (302): 157-175.
- Podolskaya ES, Anders KH, Haunert JH, Sester M (2007) Quality assessment for polygon generalization. *Proceedings of the 5th International Symposium on Spatial Data Quality*, Enschede, The Netherlands.
- Sester M (2008) Self-Organizing Maps for density-preserving reduction of objects in cartographic generalization. In *Self-Organizing Maps Applications in Geo-*

- graphic Information Science, edited by Agarwal P, Skupin A. England: John Wiley & Sons, 107-120.
- Stanislawski, LV, Savino S (2011) Pruning of hydrographic networks: A comparison of two approaches. 14th ICA/ISPRS Workshop on Generalisation and Multiple Representation, 2011, Paris.
- Töpfer F, Pillewiser W (1966) The principles of selection. *Cartographic Journal*. 3(1): 10-16.
- Wolf GW (1988) Weighted surface networks and their application to cartographic generalization. In *Visualization Technology and Algorithm*, edited by Barth W. Berlin: Springer-Verlag, 199–212.
- Zhou S, Jones JB (2005) Shape-Aware Line Generalisation with Weighted Effective Area. *Developments in Spatial Data Handling*. Springer, 369-380.